

# CLUSTERING HYPERTEXT WITH APPLICATIONS TO WEB SEARCHING

## BACKGROUND OF THE INVENTION

### *Field of the Invention*

5           The present invention generally relates to search programs and more particularly to an improved method and system for searching which clusters hypertext documents.

### *Description of the Related Art*

10           The World-Wide-Web has attained a gargantuan size (Lawrence, S., and Giles, C.L. Searching the World Wide Web. Science 280, 5360 (1998), 98., incorporated herein by reference) and a central place in the information economy of today. Hypertext is the *lingua franca* of the web. Moreover, scientific literature, patents, and law cases may be thought of as logically hyperlinked. Consequently, searching and organizing unstructured collections of hypertext  
15 documents is a major contemporary scientific and technological challenge.

          Given a "broad-topic Query" (Kleinberg, J. Authoritative sources in a hyperlinked environment, in *ACM-SIAM SODA* (1998), incorporated herein by reference), a typical search engine may return a large number of relevant

documents. Without effective summarization, it is a hopeless and enervating task to sort through all the returned documents in search of high-quality, representative information resources. Therefore, there is a need for an automated system that summarizes the large volume of hypertext documents returned during internet searches.

## SUMMARY OF THE INVENTION

It is, therefore, an object of the present invention to provide a structure and method for searching a database of documents comprising performing a search of the database using a query to produce query result documents, constructing a word dictionary of words within the query result documents, pruning function words from the word dictionary, forming first vectors for words remaining in a word dictionary, constructing an out-link dictionary of documents within the database that are pointed to by the query result documents, adding the query result documents to the out-link dictionary, pruning documents from the out-link dictionary that are pointed to by fewer than a first predetermined number of the query result documents, forming second vectors for documents remaining in the out-link dictionary, constructing an in-link dictionary of documents within the database that point to the query result documents, adding the query result documents to the in-link dictionary, pruning documents from the in-link

dictionary that point to fewer than a second predetermined number of the query result documents, forming third vectors for documents remaining in the in-link dictionary, normalizing the first vectors, the second vectors, and the third vectors to create vector triplets for document remaining in the in-link dictionary and the out-link dictionary, clustering the vector triplets using the following four step process of the *toric k-means* process:

- (a) arbitrarily segregate the vector triplets into clusters,
- (b) for each cluster, computing a set of concept triplets describing the cluster,
- (c) re-segregate the vector triplets into more coherent set of clusters obtained by putting each vector triplet into the cluster corresponding to the concept triplet that is closest to, that is, most similar to, the given vector triplet,
- (d) repeating steps (b)--(c) until coherence of the obtained clusters no longer significantly increases, and the process concludes by annotating the clusters using nuggets of information, the nuggets including *summary*, *breakthrough*, *review*, *keyword*, *citation*, and *reference*.

The summary comprises a document in a cluster having a most typical in-link feature vector amongst all the documents in the cluster. The breakthrough comprises a document in a cluster having a most typical in-link feature vector amongst all the documents in the cluster. The review comprises a document in a cluster having a most typical out-link feature vector amongst all the documents in

the cluster. The keyword comprises a word in a word dictionary for the cluster that has the largest weight. The citation comprises a document in a cluster representing a most typical in-link into a cluster. The reference comprises a document in a cluster representing a most typical out-link out of a cluster.

5

## **BRIEF DESCRIPTION OF THE DRAWINGS**

The foregoing and other objects, aspects and advantages will be better understood from the following detailed description of a preferred embodiment of the invention with reference to the drawings, in which:

10

Figure 1 is a schematic diagram illustrating the clustered sets of hypertext documents;

Figure 2 is a chart showing the relationship between hypertext documents;

Figure 3 is a chart showing the relationship between various search terms;

Figure 4 is a chart showing the results of various search terms;

15

Figure 5 is a schematic diagram illustrating the relationship between words, in-links, and out-links;

Figure 6 is a chart illustrating various parameter tuples;

Figure 7 is a chart illustrating results of a query;

Figure 8 is a chart illustrating the results of various search terms;

Figure 9 is a schematic diagram illustrating a hardware embodiment of the invention; and

Figure 10 is a flowchart representation of one embodiment of the invention.

## **DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS OF THE INVENTION**

As mentioned above, there is a need for an automated system that summarizes the large volume of hypertext documents returned during internet searches. The invention clusters the set of hypertext documents that are returned by a search engine in response to a broad-topic query into various clusters such that documents within each cluster are "similar" to each other. Clustering provides a way to organize a large collection of unstructured, unlabeled hypertext documents into labeled categories that are discriminated and disambiguated from each other. The invention captures the gist of each cluster using a scheme for cluster annotation that provides useful starting points for navigating/surfing in and around each cluster.

Clustering separates unrelated documents and groups of related documents, and is useful for discrimination, disambiguation, summarization, organization, and navigation of unstructured collections of hypertext documents.

The invention proposes a novel clustering process that clusters hypertext documents using words (contained in the document), out-links (from the document), and in-links (to the document). The invention automatically determines the relative importance of words, out-links, and in-links for a given collection of hypertext documents. The invention annotates each cluster using six information nuggets: *summary*, *breakthrough*, *review*, *keywords*, *citation*, and *reference*. These nuggets constitute high-quality information resources, and are extremely effective in compactly summarizing and navigating the collection of hypertext documents. The following description employs web searching as an application to illustrate its results and, as would be known by one ordinarily skilled in the art given this disclosure, the invention is not limited to web searching but instead, is equally applicable to any form of computerized searching.

Ignoring the semantic information present in various hypertext markup language (HTML) tags, a hypertext document has three different features: (i) the words contained in the document, (ii) out-links, that is, the list of hypertext documents that are *pointed to* or *cited by* the document, and (iii) the in-links, that is, the list of hypertext documents that *point to* or *cite the* document. The invention exploits all three features to cluster a collection of hypertext documents. If two documents share one or more words, the invention considers them to be semantically similar.

Extending this notion to links, if two documents share one or more out-links or in-links, the invention considers them to be similar as well. This observation is an important feature of the invention. The invention proposes a precise notion to capture the similarity between two hypertext documents along all the three features in a unified fashion. By exploiting the invention's new similarity measure, the invention proposes a geometric hypertext clustering process: *the toric k-means* that extends the classical Euclidean *k-means* process (Hartigan, J.A. *Clustering Algorithms* Wiley, 1975, incorporated herein by reference) and the spherical *k-means* process (Rasmussen, E. *Clustering Algorithms In Information Retrieval: Data Structures and Processs* (1992), W. B. Frakes and R. Baeza-Yates, Eds., Prentice Hall, Englewood Cliffs, New Jersey, pp. 419-442 Dhillon, I. S., and Modha, D. S. Concept decompositions for large sparse text data using clustering Tech. Rep. Research Report RJ 10147 (95022), IBM Almaden Research Center, July 8, 1999, incorporated herein by reference).

The invention annotates each cluster generated by the toric *k-means* process using six information nuggets: *summary, breakthrough, review, keywords, citation, and reference*. The summary and the keywords are derived from words, the review and the references are derived from out-links, and the breakthrough and the citations are derived from in-links. These nuggets constitute high-quality, typical information resources and are extremely effective in compactly summarizing and navigating the collection of hypertext documents.

The relative importances of the words, the out-links, and the in-links are tunable parameters in the inventions process. The invention proposes an *adaptive* or *data-driven* scheme to determine these parameters with the goal of simultaneously improving the quality of all the six information nuggets for each cluster.

The invention employs web searching as an application to illustrate the inventions results. Anecdotally, when applied to the documents returned by an internet search engine in responses to the queries *latex*, *abduction*, *guinea*, and *abortion*, the invention's process separates documents about "latex allergies" from those about "TeX&LaTeX" separates documents about "alien abduction" from those about "child abduction," separates documents about "Papua New Guinea," "Guinea Bissau," and "Guinea pigs" from each other, and separates documents about "pro-life" from those about "pro-choice," respectively.

## A GEOMETRIC ENCODING OF THE WEB

**The Data Set.** Suppose the invention is given a collection of hypertext documents, say,  $W$ . Let  $Q$  denote a subset of  $W$ . For example,  $W$  denotes the entire web, and  $Q$  denotes a small collection of hypertext documents retrieved by a conventional search engine (e.g., AltaVista-www.altavista.com) in response to a query. The invention is interested in clustering the hypertext documents in  $Q$ .

The situation of interest is depicted in Figure 1, which only shows those



documents that are at most one out- or in-link away from the documents in  $Q$ ; all other link information is discarded. The words contained in hypertext documents are not shown in Figure 1. The documents  $\{A, C, M, N\}$  are not in  $Q$ , but are hyperlinked to the documents in  $Q$ .

5           The invention extracts useful features from  $Q$  and proposes a geometric representation for these features. The invention will represent each hypertext document in  $Q$  as a triplet of unit vectors  $(D, F, B)$ . These component vectors are to be thought of as column vectors. The components  $D$ ,  $F$ , and  $B$  will capture the information represented by the words contained in the document, the out-links  
10           originating at the document, and the in-links terminating at the document, respectively. The method of computing these triplets for each document in  $Q$  is shown below.

**Words.** The creation of the first component  $D$  is a standard exercise in text mining or information retrieval (see Salton, G., and McGill, M. J.  
15           *Introduction to Modern Retrieval* McGraw-Hill Book Company, 1983, incorporated herein by reference).

The basic idea is to construct a *word dictionary* of all the words that appear in any of the documents in  $Q$ , and to prune or eliminate "function" words from this dictionary that do not help in semantically discriminating one cluster  
20           from another. The invention eliminated those words which appeared in less than 2 documents, standard stopwords (Frakes, W. B., and Baeza-Yates, R.

*Information Retrieval: Data Structures and Processss* Prentice Hall, Englewood Cliffs, New Jersey, 1992, incorporated herein by reference), and the HTML tags.

Suppose  $d$  unique words remain in the dictionary after such elimination.

The invention assigns a unique identifier from 1 to  $d$  to each of these words.

5 Now, for each document  $x$  in  $Q$ , the first vector  $D$  in the triplet will be a  $d$ -dimensional vector. The  $j$ th column entry,  $1 \leq j \leq d$ , of  $D$  is the number of occurrences of the  $j$ th word in the document  $x$ .

**Out-links.** The invention now outlines the creation of the second component  $F$ . The basic idea is to construct an *out-link dictionary* of all the  
10 hypertext documents in  $W$  that are pointed to by any of the documents in  $Q$ . The invention also adds each document in  $Q$  to the out-link dictionary. For example, in Figure 1, the out-link dictionary is  $\{E, G, H, I, J, K, L, M, N\}$ .

To treat nodes in  $W \setminus Q$  and in  $Q$  in a uniform fashion, the invention adds a self-loop from every document in  $Q$  to itself. Any document in the out-link  
15 dictionary that is not pointed to by at least two documents in  $Q$  provides no *discriminating* information. Hence, the invention prunes or eliminates all documents from the out-link dictionary that are pointed to by fewer than two documents (also counting the self-loops) in  $Q$ . For example, in Figure 1, the invention eliminates the node  $N$  as it is pointed to by only  $L$ , but retain  $M$  as it is  
20 pointed to by both  $G$  and  $I$ . Similarly, the invention eliminates the nodes  $E, H, K$ , and  $L$  as they are not pointed to by any document in  $Q$  other than themselves, but

retain  $G$ ,  $I$ , and  $J$  as they are pointed to by at least one other document in  $Q$  and by themselves.

Suppose  $f$  unique nodes remain in the dictionary after such elimination.

The invention assigns a unique identifier from 1 to  $f$  to each of these documents.

5 Now, for each document  $x$  in  $Q$ , the second vector  $F$  in the triplet will be a  $f$ -dimensional vector. The  $j$ th column entry,  $1 \leq j \leq f$ , of  $F$  is the number of links to the  $j$ th retained node from the document  $x$ . Figure 2 shows the out-link feature vectors for the example in Figure 1.

**In-links.** The creation of  $B$  is similar to that of  $F$ ; for completeness, the  
10 following briefly describes its construction. The basic idea is to construct an *in-link dictionary* of all the hypertext documents that point to any of the documents in  $Q$ . The invention also adds each document in  $Q$  to the in-link dictionary.

To treat nodes in  $W \setminus Q$  and in  $Q$  in a uniform fashion, the invention adds a  
15 self-loop from every document in  $Q$  to itself. Any document in the in-link dictionary that does not point to at least two documents in  $Q$  provides no *discriminating* information. Hence, the invention prunes or eliminates all documents from the in-link dictionary that point to fewer than two documents (also counting the self-loops) in  $Q$ .

20 Suppose  $b$  unique nodes remain in the dictionary after such elimination. Assign an unique identifier from 1 to  $b$  to each of these documents. Now, for

each document  $x$  as in  $\mathcal{Q}$ , the third vector  $B$  in the triplet will be a  $b$ -dimensional vector. The  $j$ th column entry,  $1 \leq j \leq b$ , of  $B$  is the number of links from the  $j$ th retained node to the document  $x$ .

**Normalization.** Finally, for each document  $x$  in  $\mathcal{Q}$ , each of the three components  $D$ ,  $F$ , and  $B$  is normalized to have a unit Euclidean norm, that is, their directions are retained and their lengths are discarded.

**Torus.** The invention now briefly points out the geometry underlying the inventions three-fold vector space models. Suppose that the invention has  $n$  documents in  $\mathcal{Q}$ . The invention denotes each document triplet as

$$x_i = (D_i, F_i, B_i), 1 \leq i \leq n.$$

Observe that, by construction, the component vectors  $D_i$ ,  $F_i$ , and  $B_i$  all have unit Euclidean norm, and, hence, can be thought of as points on the unit spheres  $S^d$ ,  $S^f$ , and  $S^b$  in dimensions  $d$ ,  $f$ , and  $b$ , respectively. Thus, each document triplet  $x_i$  lies on the product space of three spheres, which is a *torus* (see [www.treasure-troves.com/math/Torus.html](http://www.treasure-troves.com/math/Torus.html)). Furthermore, by construction, the individual entries of the component vectors  $D_i$ ,  $F_i$ , and  $B_i$  are nonnegative, hence, the component vectors are in fact in the nonnegative orthants of  $R^d$ ,  $R^f$ , and  $R^b$ , respectively. For notational convenience, the invention refers to the intersection of  $(S^d \times S^f \times S^b)$  with the non-negative orthant of  $R^{d+f+b}$  as  $T$ .

**Conventional internet search engine.** The invention runs a user query through a conventional internet search engine which typically returns a list of 200 URLs containing the keywords in the query. The invention crawls and retrieves each of these 200 documents (for convenience, those documents that could not be  
5 retrieved in 1 minute were discarded), and that becomes the invention's query result document set  $Q$ . Next, the invention parses each of these documents and constructs the unpruned out-link dictionary. Finally, for each document in  $Q$ , using queries of the form "link:URL" on the conventional internet search engine, the invention retrieves the URLs of the top 20 documents that point to it. This  
10 constitutes the inventions unpruned in-link dictionary. Observe that the invention does not need the actual documents in either the out- or the in-link dictionary. The set  $Q$  and the out- and the in-link dictionaries now become the inputs for the vector space model construction procedure described above.

**Statistics.** In Figure 3, for a number of queries, the invention presents  
15 statistical properties of the three-fold vector space models. A note of notation:  $n$  represent the number of documents in  $Q$ ,  $d^o$  and  $d$  are the number of words in the word-dictionary before and after elimination of function words, respectively,  $d^o$  is the average number of nonzero word counts per document, and  $n_d$  is the number of documents which contain at least one word after elimination. The symbols  $f^o$ ,  $f$ ,  
20  $f^o$ , and  $n_f$  as well as the symbols  $b^o$ ,  $b$ ,  $b^o$ , and  $n_b$  have similar meaning to their counterpart for the words. The symbols  $n_f$  and  $n_b$  are the number of documents in

$Q$  that are eventually retained in the final, pruned out-link and the in-link dictionaries, respectively.

**High-dimensional.** By observing the  $d, f$ , and  $b$  values in Figure 3, the invention observes that, even after pruning, the word, out-link, and in-link dictionaries are found to be very high-dimensional. Also, typically,  $d$  is the much larger than both  $f$  and  $b$ .

**Sparse.** By observing the ratios  $d^o/d, f^o/f$ , and  $b^o/b$  in Table 1, the invention observes that the vector space models are very sparse. A sparsity of 96% is typical for words, that is, on an average each document contains only 4% of the words in the word dictionary. Similarly, sparsities of 95-98% and 91-97% are typical for out- and in-links, respectively.

By observing the  $n_f$  and  $n_b$  values, the invention observes that not all documents have nonzero out-link and in-link features vectors. This points once again to the sparse link topology that is holding the web together. Also, the variations in  $n_f$  and  $n_b$  values point to the fact that some "topics" or "communities" are more tightly coupled than others.

**Value of  $W \setminus Q$ .** Finally, by observing the  $\hat{n}_f$  and  $\hat{n}_b$  value, the invention observes that the number of nodes from the original set  $Q$  retained in the final out-link and in-link dictionaries (after pruning) is rather small. In other words, the interconnection structure of the set  $Q$  with the rest of the web, namely,  $W \setminus Q$ , contains the vast majority of the link information in the invention's feature

vectors. This justifies the invention's inclusion of links between the documents in  $Q$  and the documents  $W \setminus Q$ .

### TORIC $k$ -MEANS PROCESS

**A Measure of Similarity.** Given document triplets  $x = (D, F, B)$  and  $\tilde{x} = (\tilde{D}, \tilde{F}, \tilde{B})$  on the torus  $T$ , the invention defines a measure of similarity between them as a weighted sum of the inner products between the individual components. Precisely, the invention writes

$$S(x, \tilde{x}) = \alpha_w D^T \tilde{D} + \alpha_f F^T \tilde{F} + \alpha_b B^T \tilde{B}, \quad (1)$$

where *weights*  $\alpha_w$ ,  $\alpha_f$  and  $\alpha_b$  are nonnegative numbers such that

$$\alpha_w + \alpha_f + \alpha_b = 1$$

Observe that for any two document triplets  $x$  and  $\tilde{x}$ ,  $0 \leq S(x, \tilde{x}) \leq 1$ . Also,

observe that if the invention sets  $\alpha_w = 1$ ,  $\alpha_f = 0$ , and  $\alpha_b = 0$ , then the invention gets the classical cosine similarity between document vectors that has been widely used in information retrieval (Salton, et al.(1983)). The parameters

$\alpha_w$ ,  $\alpha_f$  and  $\alpha_b$  are tunable in the inventions process to assign different weights

to words, outlinks, and in-links as desired. The invention will later illustrate, in detail, the appropriate choice of these parameters.

**Concept Triplets.** Suppose the invention is given  $n$  document vector

triplets  $x_1, x_2, \dots, x_n$  on the torus  $T$ . Let  $\pi_1, \pi_2, \dots, \pi_k$  denote a partitioning of these document triples into  $k$  disjoint clusters. For each fixed  $1 \leq j \leq k$ , the *concept vector triplet* or concept triplet, for short, is defined as

$$c_j = (D_j^*, F_j^*, B_j^*)$$

(3)

where  $x = (D, F, B)$  and  $\|\cdot\|$  denotes the Euclidean norm. Observe that, by construction, each component of the concept triplet has unit Euclidean norm. The concept triplet  $c_j$  has the following important property. For any triplet

$\tilde{x} = (\tilde{D}, \tilde{F}, \tilde{B})$  on the torus  $T$ , the invention has from the Cauchy-Schwarz

inequality that

$$\sum_{x \in \pi_j} S(x, \tilde{x}) \leq \sum_{x \in \pi_j} S(x, c_j) \quad (4)$$

In Figure 3,  $n$  represents the number of documents in  $Q$ ,  $d^\circ$  and  $d$  are the number of words in the word-dictionary before and after elimination of function words, respectively,  $d^\circ$  is the average number of nonzero word counts per document, and  $nd$  is the number of documents which contain at least one word after elimination. The symbols  $f^\circ, f, f^\circ$ , and  $n_f$  as well as the symbols  $b^\circ, b, b^\circ$ , and  $n_b$  have a similar meaning to their counterparts for the words. The symbols  $n_f$  and



$n_i$  are the number of documents in  $Q$  that are eventually retained in the final, pruned out-link and the in-link dictionaries, respectively.

Thus, in an average sense, the concept triplet may be thought of as being the closest in  $S$  to all the document vector triplets in the cluster  $\pi_j$ . The invention demonstrates that concept triplets contain valuable conceptual or semantic information about the clusters that is important in interpretation and annotation.

**The Objective Function.** Motivated by (4), the invention measures the "coherence" or "quality" of each cluster  $\pi_j, 1 \leq j \leq k$ , as

$$\sum_{x \in \pi_j} S(x, c_j)$$

If all documents in a cluster are identical, then the average coherence of that cluster will have the highest possible value of 1, while if the document vectors in a cluster vary widely, then the average coherence will be small, that is, close to 0.

The invention measures the quality of any given partitioning  $\pi_j, 1 \leq j \leq k$  using the following *objective function*:

$$\sum_{j=1}^k \sum_{x \in \pi_j} S(x, c_j) \quad (5)$$

Intuitively, the objective function measures the combined coherence of all the  $k$  clusters.

**The Process.** The invention's objective is to find  $k$  disjoint clusters clusters  $\pi_1^\dagger, \pi_2^\dagger, \dots, \pi_k^\dagger$  such that the following is maximized

(6)

Even when only one of the parameters  $\alpha_d$ ,  $\alpha_f$ , or  $\alpha_b$  is nonzero, finding the optimal solution to the above maximization problem is known to be NP-complete.

Discussed below is an efficient and effective approximation process: the *toric k-means* that may be thought of as a *gradient ascent* method.

**Step 1.** Start with an arbitrary partitioning of the document vectors,

namely,  $\{\pi_j^{(0)}\}_{j=1}^k$ . Let  $\{c_j^{(0)}\}_{j=1}^k$  denote the concept triplets associated with the

given partitioning. Set the index of iteration  $t = 0$ . The choice of the initial

partitioning is quite crucial to finding a "good" local minima; for recent work on this area, see (Bradley, P., and Fayyad, U. Refining initial points for k-means clustering In *ICML* (1998), pp. 91-99, incorporated herein by reference).

**Step 2.** For each document vector triplet  $x_i$ ,  $1 \leq i \leq n$  the invention finds the concept triplet that is closest to  $x_i$ . Now, for  $1 \leq j \leq k$ , compute the new partitioning  $x_i, 1 \leq i \leq n$

induced by the old concept triplets  $\{c_j^{(t)}\}_{j=1}^k$ :

$$\pi_j^{(t+1)} = \left\{ x \in \{x_i\}_{i=1}^n : S(x, c_j^{(t)}) \geq S(x, c_\ell^{(t)}), 1 \leq \ell \leq k \right\} \quad (7)$$

In words,  $\pi_j^{(t+1)}$  is the set of all document vector triplets that are closest to the concept triplet  $c_j^{(t)}$ . If it happens that some document triplet is simultaneously closest to more than one concept triplet, then it is randomly assigned to one of the clusters.

**Step 3.** Compute the new concept triplets  $\{c_j^{(t+1)}\}_{j=1}^k$  corresponding to

the partitioning computed in equation (7) by using (2)-(3) where instead of  $\pi_j$

The invention uses  $\pi_j^{(t+1)}$ .

**Step 4.** If some "stopping criterion" is met, then set  $\pi_j^\dagger = \pi_j^{(t+1)}$  and set

$c_j^\dagger = c_j^{(t+1)}$  for  $1 \leq j \leq k$ , and exit. Otherwise, increment  $t$  by 1, and go to step 2

above. An example of a stopping criterion is: Stop if the change in the objective function, between two successive iterations, is less than some specified threshold.

**Shape of Clusters.** Clusters defined using equation (7) are known as

*Voronoi* or *Dirichlet* partitions. The boundary between two clusters, say,  $\pi_j^\dagger$  and  $\pi_\ell^\dagger$ , is the locus of all document triplets  $x$  on  $T$  satisfying:

$$S(x, c_j^\dagger) = S(x, c_\ell^\dagger).$$

5 If only one of the parameters  $\alpha_d$ ,  $\alpha_f$ , or  $\alpha_b$  is nonzero, then the above locus is a hypercircle on the corresponding sphere; when more than one parameters is nonzero, the locus is a hyperellipsoid. Thus, each cluster is a region on the surface of the underlying torus is bounded by hyperellipsoids. In conclusion, the geometry of the torus plays an integral role in determining the "shape" and the  
10 "structure" of the clusters found by the toric  $k$ -means process.

#### CLUSTER ANNOTATION AND INTERPRETATION.

Suppose the invention has clustered a hypertext collection  $Q$  into  $k$  clusters

$\{\pi_j^\dagger\}_{j=1}^k$ ; let  $\{c_j^\dagger\}_{j=1}^k$  denote the corresponding concept triplets. In this raw form,

the clustering is of little use. The invention now uses the concept triplets to  
15 interpret and annotate each cluster. The process of seeking good cluster annotation will motivate the choice of the weights  $\alpha_d$ ,  $\alpha_f$ , or  $\alpha_b$ .

The invention fixes a cluster  $\pi_j^\dagger, 1 \leq j \leq k$ . Let  $c_j^\dagger = (D_j^*, F_j^*, B_j^*)$  denote the corresponding concept triplet. The following shows how to label the fixed cluster  $\pi_j^\dagger$  using six different nuggets of information.

**Summary.** A *summary* is a document in  $\pi_j^\dagger \subset Q$  that has the most typical in-link feature vector amongst all the documents in the cluster. Formally, the summary is a document triplet  $x = (D, F, B)$  whose word component  $D$  is closest in cosine similarity to  $D_j^*$ .

**Breakthrough.** A *breakthrough* is a document in  $\pi_j^\dagger \subset Q$  that has the most typical in-link feature vector amongst all the documents in the cluster. Formally, the breakthrough is a document triplet  $x = (D, F, B)$  whose in-link component  $B$  is closest in cosine similarity to  $B_j^*$ .

**Review.** A *review* is a document in  $\pi_j^\dagger \subset Q$  that has the most typical out-link feature vector amongst all the documents in the cluster. Formally, the review is a document triplet  $x = (D, F, B)$  whose out-link component  $F$  is closest in cosine similarity to  $F_j^*$ .

**Keywords.** *Keywords* for the cluster  $\pi_j^\dagger$  are those words in the word dictionary that have the largest weight in  $D_j^*$  compared to their respective weights

in  $D_\ell^*$ ,  $1 \leq \ell \leq k, \ell \neq j$ . Keywords are the most discriminating words in a cluster and constitute an easy-to-interpret cluster signature.

**Citations.** *Citations* for the cluster  $\pi_j^\dagger$  are those in-links in the in-link dictionary that have the largest weight in  $B_j^*$  compared to their respective weights in  $B_\ell^*$ ,  $1 \leq \ell \leq k, \ell \neq j$ . Citations represent the set of most typical links entering (the documents in) the given cluster.

**References.** *References* for the cluster  $\pi_j^\dagger$  are those out-links in the out-link dictionary that have the largest weight in  $F_j^*$  compared to their respective weights in  $F_\ell^*$ ,  $1 \leq \ell \leq k, \ell \neq j$ . References represent the set of most typical links exiting (from the documents in) the given cluster.

If the invention were to cluster a collection of not-hyperlinked text documents, then the summary and the keywords would constitute an adequate annotation. For hypertext collections, the inventions annotation naturally extends the concepts of summary and the keywords from words to in-links and out-links as well. The names of the six nuggets have been inspired by their respective analogues in the scientific literature. Observe that the summary, the breakthrough, and the review are meant to be primarily *descriptive* of the contents of the cluster, whereas the keywords, the references, and the citations are meant to

be *discriminative* characteristics of the cluster. Also, observe that the summary, the breakthrough, and the review are, by definition, drawn from the set  $Q$ ; however, the citations and the references may or may not be in the set  $Q$ .

**Effectiveness of Annotation: Examples.** Suppose, for a moment, that the invention is not interested in clustering at all; in other words, suppose that the invention is interested in only one cluster, that is,  $k = 1$ . Even in this case, the six nuggets described above are meaningful, and often capture the top information resources present in  $Q$ .

For example, in Figure 4, by treating the entire set  $Q$  as one cluster, the invention presents the six nuggets for each of the four queries: *virus*, "*human rights*," *dilbert*, and *terrorism*. As even a casual glance reveals, the annotation indeed captures the top information resources in every case and provides a valuable starting point for navigating the documents surrounding the cluster. By treating the entire set  $Q$  as one cluster, we present the corresponding six nuggets for each of the four queries: *virus*, "*human rights*", *dilbert*, and *terrorism*. Every document that is in  $Q$  is followed by a number in the parentheses that represents its rank in the documents returned by AltaVista. Every summary, review, and breakthrough is always followed by a parenthetic number, whereas the references or citations are followed by a parenthetic number only when applicable.

Furthermore, note that, in Figure 4, every document that is in  $Q$  is followed by parenthetic number that represents its rank in the documents returned

by conventional internet search engine. For example, for the query "virus" the summary is "Anti-Virus Tools (51)" meaning that it was the fifty-first document returned by a conventional internet search engine. By observing these parenthetic numbers, the invention can conclude that, in almost every case, the top resources found by the invention's annotation were not amongst the top documents returned by the conventional internet search engine. For example, for the query "human rights," the inventions annotation finds the "United Nations Human Rights Website" as a breakthrough, while it is the twenty-second document returned by conventional internet search engine. Thus, in its simplest form, the inventions annotation provides a rearrangement of the results returned by the conventional internet search engines. Such rearrangements are important, since user studies have shown that the users rarely go beyond the top 20 documents returned by an information retrieval system (Silverstein, C., Henzinger, M., Marais, J., and Moricz, M. Analysis of a very large AltaVista query log. Tech. Rep. 1998-014, Compaq Systems Research Center, Palo Alto, CA, October 1998., incorporated herein by reference).

## **CHOICE OF THE WEIGHTS**

In the end, it is really the annotation of each cluster in terms of the above six nuggets that is presented to the end user.



As shown in Figure 4, by treating the entire set  $Q$  as one cluster, the invention presents the corresponding six nuggets for each of the four Queries: *virus*, *"human rights"*, *dilbert*, and *terrorism*. Every document that is in  $Q$  is followed by a number in the parentheses that represents its rank in the documents returned by conventional internet search engine. Every summary, review, and breakthrough is always followed by a parenthetic number, whereas the references or citations are followed by a parenthetic number only when applicable.

Hence, arguably, a natural goal of hypertext clustering is to obtain the most descriptive and discriminative nuggets possible. Clearly, if the invention uses  $\alpha_d = 1$ ,  $\alpha_r = 0$ , and  $\alpha_b = 0$ , then the invention gets a good discrimination amongst the resulting clusters in the feature space constituted by the words. Consequently, the invention obtains good summary and keywords for the resulting clusters. Similarly, if the invention uses  $\alpha_d = 0$ ,  $\alpha_r = 1$ , and  $\alpha_b = 0$ , then the invention can obtain good review and references for the resulting clusters. Finally, if the invention uses  $\alpha_d = 0$ ,  $\alpha_r = 0$ , and  $\alpha_b = 1$ , then the invention can obtain good breakthrough and citations for the resulting clusters. To truly and completely exploit the hypertext nature of the given document collection, the invention would like all the six nuggets to be of good quality *simultaneously*. This can be achieved by judiciously selecting the parameters  $\alpha_d$ ,  $\alpha_r$ , and  $\alpha_b$ . The invention now provides a formal framework for this choice.

Throughout this section, the invention fixes the number of clusters  $k \geq 2$ .

As before, let  $\alpha_d$ ,  $\alpha_f$ , and  $\alpha_b$  be nonnegative numbers that sum to 1.

Geometrically, these parameters lie on a planar triangular region, say,  $\Delta_0$ , that is shown in Figure 5. For brevity, the invention writes  $\alpha = (\alpha_d, \alpha_f, \alpha_b)$ . Let  $\Pi(\alpha) =$

5  $\left\{ \pi_j^i \right\}_{j=1}^k$ , denote the partitioning obtained by running the toric  $k$ -means process

with the parameter values  $\alpha_d$ ,  $\alpha_f$ , and  $\alpha_b$ . From the set of all possible clusterings

$\left\{ \Pi(a) : a \in \Delta_0 \right\}$ . The invention selects a partitioning that yields the *best* cluster

annotations. Towards this goal, the invention introduces a figure-of-merit for evaluating and comparing various clusterings.

10 **Fixing a clustering  $\Pi(\alpha)$ .** For the given clustering, the summary, which is a descriptive characteristic, for each of the clusters will be good if each cluster is as coherent as possible in the word feature space, that is, if the following is maximized:

where  $x = (D, F, B)$ . Furthermore, the keywords, which are a discriminative

15 characteristics, will be good if the following is minimized:

where  $x = (D, F, B)$ . Intuitively,  $\Gamma_d(\alpha)$  and  $\Lambda_d(\alpha)$  capture the *average within cluster coherence* and *average between cluster coherence*, respectively, of the

clustering  $\Pi(\alpha)$  in the word feature space. The summary and the keywords both will be good if the following ratio is maximized:

(8)

where  $n_d$  denotes the number of document triplets in  $Q$  that have a non-zero word feature vector; see, for example, Figure 3. In the case that  $\Lambda_d(\alpha) = 0$ , the clusters are *perfectly separated* in the word feature space.

The quantities  $\Gamma_r(\alpha)$ ,  $\Lambda_r(\alpha)$ ,  $\Gamma_b(\alpha)$ ,  $\Lambda_b(\alpha)$ ,  $Q_r(\alpha)$ , and  $Q_b(\alpha)$  are defined in a similar fashion. The quantity  $Q_r(\alpha)$  should be maximized to obtain good quality review and references, and the Quantity  $Q_b(\alpha)$  should be maximized to obtain good quality breakthrough and citations.

In Figure 5 the triangular region  $\Delta_0$  is formed by the intersection of the plane  $\alpha_d + \alpha_f + \alpha_b = 1$  with the nonnegative orthant of  $R^3$ . The left-vertex, the right-vertex, and the top-vertex of the triangle correspond to the points  $(1, 0, 0)$ ,  $(0, 1, 0)$ , and  $(0, 0, 1)$ , respectively.

The following shows how the invention selects the optimal parameter tuple  $a^*$  and the corresponding clustering  $\Pi(a^*)$ .

**Step 1.** Theoretically, the invention would like to run the tortic k-means process for every parameter triplet in:

$$\Delta_0 = \{\alpha : \alpha_d + \alpha_f + \alpha_b = 1, \alpha_d, \alpha_f, \alpha_b \geq 0\}. \quad (9)$$

In practice, the invention replaces the region  $\Delta_0$  in equation (9) by a finite number of points on a discrete grid that are graphically shown using the symbol in Figure 5.

**Step 2.** To obtain good cluster annotations in terms of all the six nuggets, the invention simultaneously maximizes  $Q_d$ ,  $Q_f$ , and  $Q_b$ . Hence, the invention selects the parameters  $a^\dagger$  as the solution of the following maximization problem:

$$a^\dagger = \arg \max_{a \in \Delta} [Q_d(a) \times Q_f(a) \times Q_b(a)] \quad (10)$$

where the invention defines the region  $\Delta$ . First, the invention needs some notation.

$$\begin{aligned} R_d &= \{a \in \Delta_0 : \Lambda_d(a) = 0\} \\ R_f &= \{a \in \Delta_0 : \Lambda_f(a) = 0\} \\ R_b &= \{a \in \Delta_0 : \Lambda_b(a) = 0\} \\ \Delta_3 &= R_d \cap R_f \cap R_b \\ \Delta_2 &= ((R_d \cap R_f) \cup (R_d \cap R_b) \cup (R_f \cap R_b)) \setminus \Delta_3 \\ \Delta_1 &= (R_d \cup R_f \cup R_b) \setminus \Delta_2 \end{aligned}$$

The invention now defines the region  $\Delta$  as follows:

$$\Delta = \begin{cases} \Delta_3 & \text{if } \Delta_3 \neq \phi \\ \Delta_2 & \text{elseif } \Delta_2 \neq \phi \\ \Delta_1 & \text{elseif } \Delta_1 \neq \phi \\ \Delta_0 & \text{otherwise.} \end{cases}$$

The following explains the reasoning behind the above definitions. The regions  $R_d$ ,  $R_f$ , and  $R_b$  denote the set of parameters for which the corresponding clusterings perfectly separate the document triplets in the word, out-link, and

in-link feature spaces, respectively. The region  $\Delta$  denotes the set of parameters for which the corresponding clusterings perfectly separate the document triplets in all the three feature spaces. Clearly, if such clusterings are available, that is, if  $\Delta$  is not empty, then the invention would prefer them. Hence, the invention sets  $\Delta = \Delta_3$ , if  $\Delta_3 \neq \emptyset$ . The region  $\Delta_2$  denotes the set of parameters for which the corresponding clusterings perfectly separate the document triplets along two, but not all three, feature spaces. In the case that  $\Delta_3$  is empty, the invention prefers clusterings in  $\Delta_2$ . Now, the region  $\Delta_1$  denotes the set of parameters for which the corresponding clusterings perfectly separate the document triplets along one and only one of the three feature spaces. In the case that  $\Delta_3$  and  $\Delta_2$  are both empty, the invention prefers the clusterings in  $\Delta_1$ . Finally,  $\Delta_0$  which is the entire triangular region in Figure 5 is the default choice when  $\Delta_3$ ,  $\Delta_2$ , and  $\Delta_1$  are all empty. In practice, the invention has found that  $\Delta_3$  and  $\Delta_2$  are usually empty, and, hence, for most data sets, the invention expects the region  $\Delta$  to be either  $\Delta_1$  or  $\Delta_0$ .

**Step 3.** Having selected the parameter triple  $a^\dagger$  let the optimum clustering  $\Pi(a^\dagger)$  denote the partitioning obtained by running the torte k-means process with  $a^\dagger$ .

To illustrate the above scheme, the invention now utilizes the  $Q_d$ ,  $Q_f$ ,  $Q_b$  and  $T = Q_d \times Q_f \times Q_b$  values for various parameter tuples, where  $Q$  is the set of documents returned by a conventional internet search engine in response to the query *guinea* and  $k = 3$ , as shown in Figure 6.

The first, second, and the third rows correspond to clustering primarily along words, out-links, and in-links, respectively, while the fourth row corresponds to the clustering corresponding to the optimal parameter tuple. It can be seen that the optimal clustering achieves significantly larger T value than clusterings which cluster only along one of the three features. In practice, the larger T value often translates into superior cluster annotation.

## RESULTS:

In Figure 7, the invention summarizes the results obtained by solving the maximization problem in equation (10) for each of the four queries: *latex*, *abduction*, *guinea*, and *abortion*. The set of documents returned by AltaVista for each of the four queries: *latex*, *abduction*, *guinea*, and *abortion* are clustered into  $k$  clusters. For each query, the optimal parameter tuple at:  $a^{\dagger}$  is determined by solving the minimization problem in (10). For queries *abduction* and *guinea*, all the three sets  $\Delta_3$ ,  $\Delta_2$ , and  $\Delta_1$  turn out to be empty, and, hence,  $\Delta = \Delta_0$ . For queries *latex* and *abortion*, the two sets  $\Delta_3$  and  $\Delta_2$  turn out to be empty, but the set  $\Delta_1$  is not empty, and, hence,  $\Delta = \Delta_1$ .

In Figure 8, the invention presents the optimal clusterings corresponding to the optimal parameter triplets in Figure 7 for the queries *latex*, *abduction*, *guinea*, and *abortion*. By running the toric  $k$ -means algorithm with the respective optimal parameter tuples in Figure 7, the set of documents  $\mathcal{Q}$  returned by

AltaVista is clustered in response to the queries *latex*, *abduction*, *guinea*, and *abortion* into  $k = 2, 2, 3$ , and 3 clusters, respectively. The six nuggets are shown for each cluster. Every document that is in  $Q$  is followed by a number in the parentheses that represents its rank in the documents returned by AltaVista. Every summary, review, and breakthrough is always followed by a parenthetic number, whereas the references or citations are followed by a parenthetic number only when applicable.

It can be seen from Figure 8 that (i) the set of documents corresponding to *latex* is neatly partitioned into "latex allergies" cluster and into " $T_E X \& L^A T_E X$ " cluster; (ii) the set of documents corresponding to *abduction* is neatly partitioned into "alien abduction" cluster and into "child abduction" cluster; (iii) the set of documents corresponding to *guinea* is neatly partitioned into "Papua New Guinea," "Guinea Bissau," and "Guinea pigs" clusters; and, finally, (iv) the set of documents corresponding to *abortion* is neatly partitioned into two "pro-life" cluster and one "pro-choice" clusters.

In Figure 10 the invention performs a search of the database 100, constructs a word dictionary 101, prunes function words 102, forms first vectors 103, constructs an out-link dictionary 104, prunes documents from the out-link dictionary 105, forms second vectors 106, constructs an in-link dictionary 107, prunes documents from the in-link dictionary 108, forms third vectors 109, normalizes the first vectors 110, arbitrarily segregates the vector triplets into

clusters 111, determines a coherence of each of the clusters 112, and re-segregates the vector triplets into fixed clusters 113.

While the overall methodology of the invention is described above, the invention can be embodied in any number of different types of systems and executed in any number of different ways, as would be known by one ordinarily skilled in the art. For example, as illustrated in Figure 9, a typical hardware configuration of an information handling/computer system in accordance with the invention preferably has at least one processor or central processing unit (CPU) 900. For example, the central processing unit 900 could include various image/texture processing units, mapping units, weighting units, classification units, clustering units, filters, adders, subtractors, comparators, etc. Alternatively, as would be known by one ordinarily skilled in the art given this disclosure, multiple specialized CPU's (or other similar individual functional units) could perform the same processing, mapping, weighting, classifying, clustering, filtering, adding, subtracting, comparing, etc.

The CPU 900 is interconnected via a system bus 901 to a random access memory (RAM) 902, read-only memory (ROM) 903, input/output (I/O) adapter 904 (for connecting peripheral devices such as disk units 905 and tape drives 906 to the bus 901), communication adapter 907 (for connecting an information handling system to a data processing network) user interface adapter 908 (for connecting a peripherals 909-911 such as a keyboard, mouse, imager, microphone,



speaker and/or other interface device to the bus 901), a printer 912, and display adapter 913 (for connecting the bus 901 to a display device 914). The invention could be implemented using the structure shown in Figure 9 by including the inventive method, described above, within a computer program stored on the storage device 905. Such a computer program would act on an image supplied through the interface units 909-911 or through the network connection 907. The system would then automatically segment the textures and output the same on the display 914, through the printer 912 or back to the network 907.

Document clustering using only textual features such as words or phrases has been extensively studied; for a detailed review of various k-means type processs, graph theoretical processs, and hierarchical agglomerative clustering processs, see Rasmussen, E. Clustering Algorithms. In Information Retrieval: Data Structures and Algorithms (1992), W.B. Frakes and R. Baeza-Yates, Eds., Prentice Hall, Englewood Cliffs, New Jersey, pp. 419-442, and Willet, P., Recent trends in hierarchic document clustering: a critical review. Inform. Proc. & Management (1988), pp. 577-597, incorporated herein by reference.

By treating the references made by one scientific paper (or a patent or a law case) to another as a logical hyperlink, one can interpret scientific literature (or patents or law cases) as a hypertext document collection. Citation analysis was developed as a tool to identify core sets or clusters of articles, authors, or journals of particular fields of study by using the logical hyperlinks between scientific

papers, see White, H.D., and McCain, K.W., *Bibliometrics*, Annual Review of Information Science and Technology 24 (1989), pp. 119-186, and Small H., Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Amer. Soc. Inform. Sci.* (1973), pp. 265-269, incorporated  
5 herein by reference. Larson, R., *Bibliometric of the world wide web: An exploratory analysis of the intellectual structure of cyberspace*, In annual Meeting Amer. Soc. Info. Sci. (1996), incorporated herein by reference, has proposed using citation analysis with multidimensional scaling to identify clusters in the web. Recently, Kleinberg, supra, has extended citation analysis to web searching. In  
10 response to a topic Query, HITS produces two distinct but interrelated types of pages: authorities (highly-cited pages) and hubs (pages that cite many authorities). HITS only uses the link topology; CLEVER refines HITS to include key word matches within anchor text, Chakrabarti, S., Dom, B. E., Raghavan, P., Rajagopalan, S., Gibson, D., and Kleinberg, J., Automatic resource compilation  
15 by analyzing hyperlink structure and associated text, In WWW7 (1998), incorporated herein by reference. For a highly accessible treatment of use of citation analysis in web searching, see Chakrabarti., S., Dom, B.E., Kumar, S. R., Raghayan, P., Rajagopalan, S., Tomkins, A., Kleinberg, J. M., and Gibson, D., Hypersearching the web., *Scientific American* (June 1999), incorporated herein by  
20 reference. The invention's method differs from citation analysis in that in

Furthermore, this body of work has not focused on hypertext clustering which is the problem of interest.

Botafogo, R. A., Cluster analysis for hypertext systems, in ACM SIGIR (1993), incorporated herein by reference, has proposed a graph-based process for clustering hypertext that uses link information but no textual information; he proposed the number of independent paths between nodes as a measure of similarity. MukherJea, S., Foley, J.D., and Hudson, S.E., Interactive clustering for navigating in hypermedia systems, In ACM Hypertext (1994), incorporated herein by reference, have proposed using content- and structure-based processs for interactive clustering of hypertext. In their model, the user precisely specifies her information need, for example, all nodes containing some content or all graphical substructures, and, hence, unlike the present invention, theirs is not an automated clustering methodology.

Weiss, R., Velez, B., Sheldon, M.A., Namprempre, C. Szilagyi, P., Duda, A., and Gifford, D.K., Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering, in ACM Hypertext (1996), incorporated herein by reference, combined information about document contents and hyperlink structures to automatically cluster hypertext documents. The invention is distinct in the choice of the processs, the underlying similarity metrics, and the cluster naming or annotation scheme. In particular, Weiss et al, supra, uses the complete link process, while the invention develops a variant of the k-means

organization of information and for browsing and searching. Their results are complementary to the inventions.

While the invention has been described in terms of preferred embodiments, those skilled in the art will recognize that the invention can be practiced with modification within the spirit and scope of the appended claims.

5

addition to the link topology, the invention also uses the entire body of the text associated with each hypertext document.

Hypertext has been used to improve information retrieval. Salton G., Associative document retrieval techniques using bibliographic information, J. ACM (1963), 440-457, incorporated herein by reference, has proposed using bibliographic information, that is, out-links or references, for improving retrieval performance. The basic idea is to extract important terms from cited documents and to add these non-local terms to the citing document. This line of investigation and its variants has been explored in Kwok, K.L., A probabilistic theory of indexing and similarity measure based on cited and citing documents, J. Amer. Soc. Inform. Sci. (1985), 342-351, Croft, W.B., and Turtle, H.R., A retrieval model for incorporating hypertext links, In ACM Hypertext (1989, Making use of hypertext links when retrieving information. In ACM European Conference on Hypertext (1992), Frei, H.P., and Steiger, D., Making use of hypertext links when retrieving information. In ACM European Conference on Hypertext (1992), and, most recently, in Chakrabarti, S., Dom, B.E., and Indyk, P., Enhanced hypertext categorization using hyperlinks, In ACM Sigmod (1998), incorporated herein by reference. The invention differs from this body of work in that the invention considers the out-links and the in-links as first-class features in their own right and does not use non-local terms from either the cited or citing documents.

process. The complete link process is Quadratic-time complexity in the number of documents, while the invention's method is linear-time complexity in the number of documents. Furthermore, their measure of similarity between two documents does not constitute a valid metric, and, hence, is not useful in a geometric

5

Previously, Pirolli, P., Pitkow, J., and Rao, R., Silk from sow's ear:

Extracting usable structures from the web, in ACM SIGCHI Human Factors

Comput. (1996), incorporated herein by reference, have combined both the link

"topology and textual similarity between items as well as usage data collected by

servers and page meta-information like title and size", Pirolli, et al., supra, did not

10

treat link topology and textual similarity differently as the invention does, but

rather represented each hypertext document as a single vector of all these features.

They left the problem of automatically categorizing hypertext documents using

their feature space to future work. Chen, C., Structuring and visualizing the www

by generalized similarity analysis, in ACM Hypertext (1997), incorporated herein

15

by reference, has proposed generalized similarity analysis that combines hypertext

linkage, content similarity, and browsing patterns or usage. Chen, C. and

Czerwinski, M., From latent semantics to spatial hypertext-An integrated

approach, in ACM Hypertext (1998), incorporated herein by reference, have

exploited generalized similarity analysis along with latent semantic indexing and

20

pathfinder network scaling to develop an integrated framework for spatial